

Day 10: Practical Social Media Data Mining

ME414: Introduction to Data Science and Big Data Analytics

LSE Methods Summer Programme

25 August 2017

Day 10 Outline

Social Media Data

Accessing social media APIs

"Web scraping"

Social Media Data

Why social media data?

- ▶ Volume and coverage
- ▶ Twitter: 328 million monthly active users, 530m tweets per day ¹
- ▶ Facebook: 1.32 billion daily active users on average for June 2017, 2 billion monthly active users as of 2017 ²
- ▶ Real time — new data is available (somewhat) publicly immediately on current events
- ▶ Metadata — geographic location, user device, profile, timestamp and other metadata is accessible.

Appeal of Social Media data

- ▶ Good case for machine learning and data mining — lots of data, lots of metadata
- ▶ Many-to-many *broadcast* text corpus
- ▶ Social network analysis: a graph of social connections

Network data structure of social media

- ▶ Broadcast
 - ▶ simplex (e.g. radio, television, smoke signal)
 - ▶ duplex (e.g. round-table meeting, walkie-talkies)
- ▶ Point-to-point: sender specifies receivers
- ▶ Social media allow many of these different forms of communication
- ▶ Twitter in particular is a completely new model of communication (social or news?)
- ▶ Every user is a sensor, receiver, and broadcaster — a distributed sensor network (Crooks et al 2012)

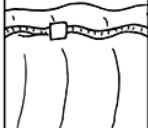
WHEN AN EARTHQUAKE HITS,
PEOPLE FLOOD THE INTERNET
WITH POSTS ABOUT IT—SOME
WITHIN 20 OR 30 SECONDS.

ROBM163 HUGE
EARTHQUAKE HERE!

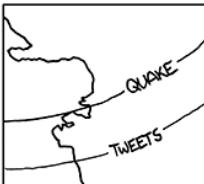


DAMAGING SEISMIC
WAVES TRAVEL AT
3-5 km/s . FIBER
SIGNALS MOVE AT
 $\sim 200,000 \text{ km/s}$.

(MINUS NETWORK LAG)



THIS MEANS WHEN THE SEISMIC
WAVES ARE ABOUT 100 KM OUT,
THEY BEGIN TO BE OVERTAKEN BY
THE WAVES OF POSTS ABOUT THEM.



PEOPLE OUTSIDE THIS RADIUS
MAY GET WORD OF THE QUAKE
VIA TWITTER, IRC, OR SMS
BEFORE THE SHAKING HITS.

WHOA!
EARTHQUAKE!



SADLY, A TWITTERER'S
FIRST INSTINCT IS NOT
TO FIND SHELTER.

RT @ROBM163 HUGE
EARTHQUAKE HERE!



Possible downsides

- ▶ Legal and ethical concerns
 - ▶ twitter is public, facebook private – see <https://twitter.com/tos?lang=en>
 - ▶ legal issues need to catch up with the technology
 - ▶ Are EULAs (End-User License Agreement) too complex to allow 'informed consent'?

Possible downsides

- ▶ Legal and ethical concerns
 - ▶ twitter is public, facebook private – see <https://twitter.com/tos?lang=en>
 - ▶ legal issues need to catch up with the technology
 - ▶ Are EULAs (End-User License Agreement) too complex to allow 'informed consent'?
- ▶ Sampling issues and many new methodological headaches

Possible downsides

- ▶ Legal and ethical concerns
 - ▶ twitter is public, facebook private – see <https://twitter.com/tos?lang=en>
 - ▶ legal issues need to catch up with the technology
 - ▶ Are EULAs (End-User License Agreement) too complex to allow 'informed consent'?
- ▶ Sampling issues and many new methodological headaches
- ▶ Biased sample (Barbera and Rivero 2013)

Possible downsides

- ▶ Legal and ethical concerns
 - ▶ twitter is public, facebook private – see <https://twitter.com/tos?lang=en>
 - ▶ legal issues need to catch up with the technology
 - ▶ Are EULAs (End-User License Agreement) too complex to allow 'informed consent'?
- ▶ Sampling issues and many new methodological headaches
- ▶ Biased sample (Barbera and Rivero 2013)
- ▶ commercial interfaces are brittle and opaque

Possible downsides

- ▶ Legal and ethical concerns
 - ▶ twitter is public, facebook private – see <https://twitter.com/tos?lang=en>
 - ▶ legal issues need to catch up with the technology
 - ▶ Are EULAs (End-User License Agreement) too complex to allow 'informed consent'?
- ▶ Sampling issues and many new methodological headaches
- ▶ Biased sample (Barbera and Rivero 2013)
- ▶ commercial interfaces are brittle and opaque
- ▶ A lot of the content is moronic

Example: Twittdiots



Michael Matthews
@YourBuddyBurns



Follow

I'm tired of this terrorist bullshit fucking w our country. Fuck it, just nuke Czechoslovakia

↩ Reply ↻ Retweet ★ Favorite ... More



InstrumentalStash
@HashHitz



Follow

I Can't believe that pair in the Boston bombing was NOT Towel heads!!! They are Czechoslovakian! Daamn!! FUCK Czechoslovakia!

↩ Reply ↻ Retweet ★ Favorite ... More



Kaitlynn Schuler
@KaitlynnSchuler



Follow

Some Czech mother fucker is about to get LITTTT up. #gethim

↩ Reply ↻ Retweet ★ Favorite ... More



s_elliott11

What did America ever do to the Czech Republic? Where even is the Czech Republic? Have fun with the devil terrorboy

🐦 2 days ago ↩ Reply ↻ Retweet ☆ Favorite



Jafar El-Shabazz
@Ilcooljaff



Follow

The media fucked up! They was sayin the suspect was a dark skinned male..turned out to be a Czech republican. ???!

↩ Reply ↻ Retweet ★ Favorite ... More

Other twitter challenges

Other twitter challenges

- ▶ Large amounts of data
 - ▶ storage problems
 - ▶ analysis problems

Other twitter challenges

- ▶ Large amounts of data
 - ▶ storage problems
 - ▶ analysis problems
- ▶ Language is informal and often non-textual (emoticons, links, images) - and slang, txtspk, emoticons :-)

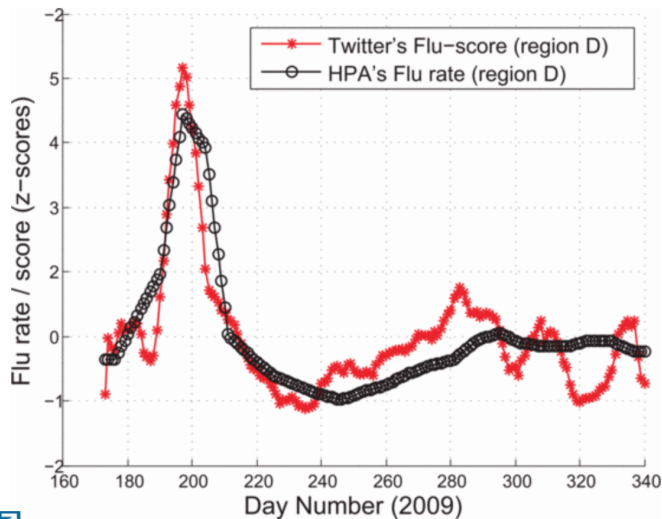
Other twitter challenges

- ▶ Large amounts of data
 - ▶ storage problems
 - ▶ analysis problems
- ▶ Language is informal and often non-textual (emoticons, links, images) - and slang, txtspk, emoticons :-)
- ▶ lots of fake users

Example applications

- ▶ Tracking disease through google search terms and social media (Lampos et al 2010)
 - ▶ Locate tweets in urban centres
 - ▶ Uses a Porter stemmer and stopwords
 - ▶ Uses regression to learn which words are associated with flu outbreaks: 97 'markers' (features)
 - ▶ Use this association to observe current outbreaks

Example applications



Example applications

- ▶ Predicting election outcomes or polls
- ▶ Sentiment: particularly for financial or corporate interests
- ▶ Government security/intelligence
- ▶ Social network analysis: a graph of social connections
- ▶ Nulty et al (2015) study of EP 2014

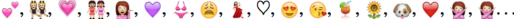

'Fixing' the biased twitter sample

Barbera, 2017, Working Paper

- ▶ Using twitter in studies of social behaviour is difficult because we lack information about the sociodemographic characteristics of twitter users
- ▶ Researchers cannot estimate 'survey' weights to recover representativeness of their samples as we do with traditional surveys
- ▶ (Additional problem: many interesting questions require demographic information!)
- ▶ Solution: match a large (250,000) sample of twitter users to voter registration records, which provide information on age, gender, race, party identification, and - indirectly - house value
- ▶ Train a classifier to learn the text features most associated with these demographics
- ▶ Predict demographics for many other users

'Fixing' the biased twitter sample

Table 4: Top predictive features (emoji, words, accounts) most associated with each category.

| | |
|--------|---|
| Female |  love, women, hair, girl, husband, mom, omg, cute, excited, <3, girls, yay, happy, hubby, boyfriend, :, can't, baby, wine, thank, heart, nails... @TheEllenShow, @khloekardashian, @MileyCyrus, @Starbucks, @jtimberlake, @VictoriasSecret, @WomensHealthMag, @channingtatum... |
| Male |  bro, man, wife, good, causewereguys, gay, great, dude, f*ck, nice, game, iphone, ni**a, church, time, #gay, girlfriend, bruh, sportscenter... @SportsCenter, @danieltosh, @MensHealthMag, @AdamScheffer, @ConanOBrien, @KingJames, @katyperry, @ActuallyNPH... |

'Fixing' the biased twitter sample

Age: 18-25

👉 🧔 🙄 😍 😭 😊 😞 😢 🔫 🩸 😟 😁 🎓 🥰 🍺 🌱

class, college, semester, life, (:, sportscenter, campus, best, literally, like, haha, just, :d, finals, classes, okay, professor, exam, studying...

@SportsCenter, @wizkhalifa, @MileyCyrus, @danieltosh, @instagram, @EmWatson, @KevinHart4real, @UberFacts, @vine...

Age: 26-40

👩🏻 🧑🏻 🦊 📺 🦵 🙄 🛩️ 🧐 💩 😁 🚲 😂 🦷 🦊 🦊

excited, work, amazing, bar, awesome, wedding, #tbt, pretty, #nofilter, ppl, bday, time, lil, #love, yay, #latergram, office, game, tonight, boo, super...

@danieltosh, @ConanOBrien, @jtimberlake, @StephenAtHome, @chelseahandler, @KimKardashian, @instagram, @NPR, @britneyspears...

Age: ≥ 40

🍰 😊 🏁 🙌 🌸 ➡ ⚾️ 💕 🌹 ✨ 🏠 ⭐ 🎸 🍷 🏈...

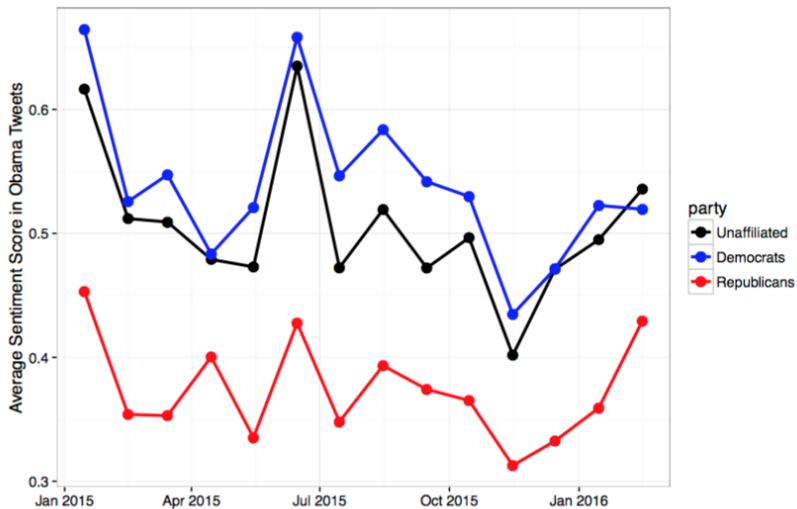
great, daughter, son, nice, r, good, ok, kids, congratulations, obama, hi, nbcthevoice, wow, happy, hope, beautiful, sorry, rock, grandson, amen. . .

@jimmyfallon, @cnnbrk, @YouTube, @Pink, @TheEllenShow, @NBCTheVoice, @SteveMartinToGo, @Oprah, @sethmeyers, @FoxNews...

'Fixing' the biased twitter sample

| | |
|--------------|--|
| Democrat |  philly, barackobama, la, sf, pittsburgh, women, nytimes, philadelphia, smh, president, gop, black, hillaryclinton, gay, republicans ... @BarackObama, @rihanna, @maddow, @billclinton, @khloekardashian, @billmaher, @Oprah, @KevinHart4real, @algore, @MichelleObama ... |
| Republican |  foxnews, #tcot, church, christmas, oklahoma, florida, obama, great, realdonaldtrump, golf, beach, megynkelly, tula, byu, seanhannity ... @FoxNews, @danieltosh, @TimTebow, @MittRomney, @taylorswift13, @jimmyfallon, @RyanSeacrest, @Starbucks, @JimGaffigan ... |
| Unaffiliated |  ohio, arkansas, columbus, cleveland, cincinnati, utah, toledo, cavs, #wps, browns, ar, akron, hogs, bengals, kent, dayton, #cbj, reds ... @instagram, @SportsCenter, @KingJames, @vine, @AnnaKendrick47, @wizkhalifa, @WhatTheFFacts, @galifianakis, @ActuallyNPH ... |

Biased sample



Social Media Data access

How can we access this data?

- ▶ API: Application Programming Interface — a way for two pieces of software to talk to each other
- ▶ Twitter, facebook, google — all expose public web services
- ▶ Your software can receive (and also send) data automatically through these services
- ▶ Data is sent by http — the same way your browser does it
- ▶ Most services have helping code (known as a wrapper) to construct http requests
- ▶ both the wrapper and the service itself are called APIs
- ▶ http service also sometimes known as REST (REpresentational State Transfer)

HyperText Transfer Protocol

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Why are we interested in HTTP?

facebook

YAHOO!

twitter

myspace.com
a place for friends

Because nearly everything a typical user does on the Internet uses HTTP

CNN.com

@mail.ru



Google
Earth

Gmail
by Google

Anatomy of a http request

```
https://api.twitter.com/1.1/search/tweets.json?  
q=Nick+Clegg%21&since_id=24012619984051000&max_id=25012619984051
```

Nick Clegg! becomes Nick+Clegg%21

- ▶ Parameters to the API are encoded in the URL
- ▶ you must encode requests — spaces and non ASCII characters are replaced

cURL and wget

- ▶ It's not usually necessary to construct these kind of requests yourself
- ▶ R, Python, and other programming languages have libraries to make it easier
- ▶ Usually you will need cURL installed to access an API, wget for downloading a website
- ▶ The documentation for the API will describe the parameters that are available.

Available social media APIs

- ▶ Wikipedia: mediawiki
- ▶ Google
 - ▶ google plus
 - ▶ blogger
- ▶ reddit
- ▶ foursquare
- ▶ facebook
- ▶ twitter: REST, Streaming, firehose, commercial

The twitter APIs: REST

- ▶ This is the most comprehensive API
- ▶ Returns a sample of historical data from the last 8–10 days.
- ▶ Stateless: you send a command and receive a result.
- ▶ http GET requests return information
- ▶ http POST requests upload or alter information (e.g. twitterbots)
- ▶ The manual: <https://dev.twitter.com/rest/public>
- ▶ R package : twitterR

The twitter APIs: Streaming

- ▶ Connect to the twitter server and collect tweets as they fly by.
- ▶ The manual: <https://dev.twitter.com/streaming/public>
- ▶ R package: streamR

Authentication

- ▶ Username and Password
- ▶ Oauth (ROauth): share a key without sharing a username and password
- ▶ IP address limitations
- ▶ Rate limitations
- ▶ Per-user and per-application

Other options

- ▶ The firehose: work with twitter
- ▶ Commercial options: GNIP (now bought by twitter) and Datasift

The Output: JSON and XML

- ▶ XML: eXtensible Markup Language: encodes documents in a form that is both human-readable and machine readable
- ▶ JSON : JavaScript Object Notation
- ▶ If you have a choice, you probably want JSON
- ▶ JSON uses key:value pairs, XML uses trees
- ▶ JSON is easily read into a programming language
- ▶ Sometimes known as serialization formats

And finally... the data.

- ▶ Full of spam, bots, unicode, and gibberish
- ▶ Lots of retweets (approximately one-third retweets, replies, tweets)
- ▶ Only 1% show location — some methods exist to infer location
- ▶ All aspects of metadata and reply/retweet structure are available
- ▶ All aspects of network structure: followers and 'friends', profile information

Twitterbots

- ▶ API also allows actions such as posting tweets (POST)
- ▶ Examples:
 - ▶ @netflix_bot posts new content using netflix api
 - ▶ @eqbot posts earthquake warnings
 - ▶ @pentametrone posts pairs of tweets in rhyming couplets ³

Twitterbots



Big Ben

@big_ben_clock



Follow

BONG BONG BONG BONG BONG BONG BONG BONG
BONG BONG

10:00 AM - 10 Oct 2014



73



63

Twitterbots in research

- ▶ Munger, 2017, Political Behaviour
- ▶ Research question: Does social sanctioning reduce racist online harassment?
- ▶ Design:
 - ▶ Randomly assign a sample of racist Twitter users to a treatment and control group
 - ▶ Treatment group: direct 'bots' representing in-group and out-group members to sanction users for their use of racist terms
 - ▶ Control group: leave users alone
 - ▶ Measure whether treatment group reduce their use of racist language in subsequent weeks

Twitterbots in research



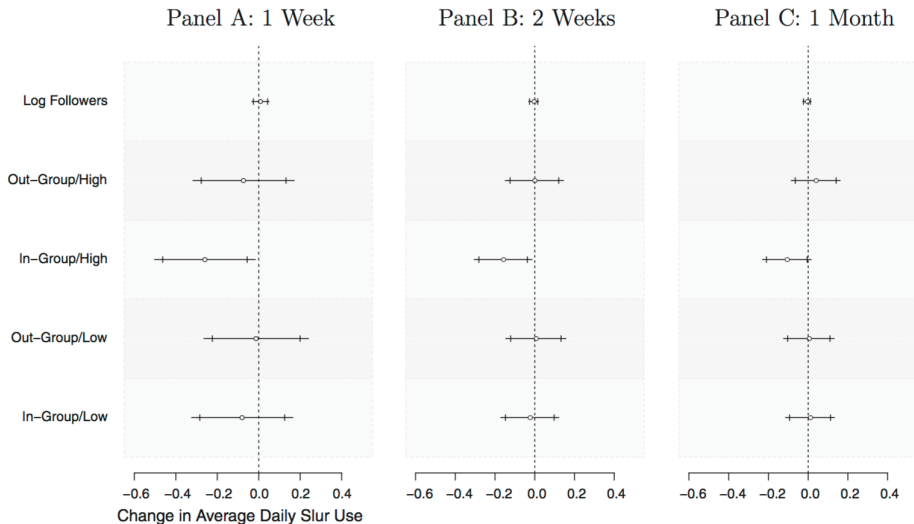
Rasheed [REDACTED]

@Rasheed [REDACTED]

@ [REDACTED] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

Twitterbots in research

Results



Twitter uses: Exploiting the meta-data (non-textual)

- ▶ location
- ▶ time
- ▶ username
- ▶ user descriptions
- ▶ networks of followers
- ▶ retweets of followers and texts

Connecting through R

R packages

- ▶ Twitter: `twitteR` for REST, `streamR` for Streaming
- ▶ Facebook: `Rfacebook`

Connecting through R

R packages

- ▶ Twitter: `twitteR` for REST, `streamR` for Streaming
- ▶ Facebook: `Rfacebook`

Python: `tweepy` and `facebook-sdk`

Connecting through R

R packages

- ▶ Twitter: twitteR for REST, streamR for Streaming
- ▶ Facebook: Rfacebook

Python: tweepy and facebook-sdk

other open-source tools exist

Connecting through R

R packages

- ▶ Twitter: `twitteR` for REST, `streamR` for Streaming
- ▶ Facebook: `Rfacebook`

Python: `tweepy` and `facebook-sdk`

other open-source tools exist

Integration with `quanteda` is fairly straightforward

Other social media access packages

- ▶ `tumblrR` R interface to the Tumblr web API
- ▶ `instaR` R interface to Instagram API
- ▶ `Rlinkedin` R interface to LinkedIn API
- ▶ `RedditExtractorR` R interface for Reddit API

Demonstration

How to get visible content directly from web pages

Scraping text from the web

- ▶ web crawlers/spider download sites by traversing links
- ▶ Python - scraPy, BeautifulSoup
- ▶ R - Rvest
- ▶ Chrome web plugins, import.io
- ▶ cUrl, wget, or other tools available ('httrack')
- ▶ Problems: rate limiting, ethical issues

Demonstration

WHenever I learn a new skill I concoct elaborate fantasy scenarios where it lets me save the day.

OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!



BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!



IT'S HOPELESS!

EVERYBODY STAND BACK.



I KNOW REGULAR EXPRESSIONS.



Make scraping unnecessary!

- ▶ Organizations and governments should be aware of need for open, machine-readable data
- ▶ data.gov.uk, data.gov
- ▶ Data should be available in human and machine format!
- ▶ Make the raw data available in as many formats as possible.
- ▶ Consider machine readability at time of data collection
- ▶ Provide an Application Programming Interface (API)